Population Estimation with Density Based Clustering

1. Introduction

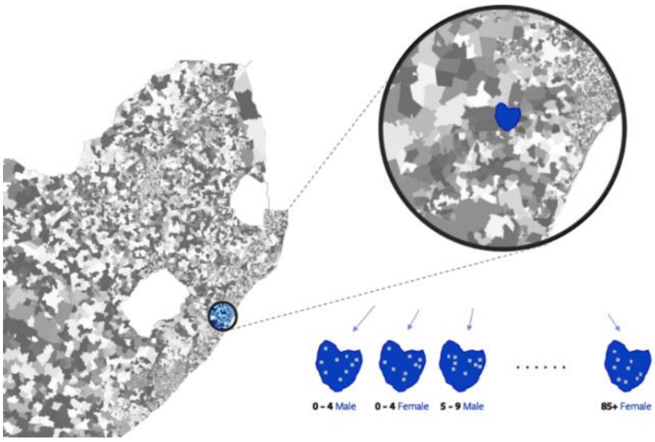
The present report focuses on a proof of concept experiment to estimate population numbers in settlements with Nord-Kivu province in Democratic Republic of Congo. The estimation process is supported by an analysis of population numbers and aims at using a Facebook Connectivity Lab high resolution dataset of population density and OSM settlements data in Nord Kivu province to determine the total number of population in a given settlement. This work is preform using Python as main coding platform.

The methods and datasets used in this proof of concept may be useful for future work, not only in DRC but also in other countries where the same kind of settlement layer created thy Facebook Connectivity Lab is available.

2. Datasets used for classification

The estimation process described here used three different datasets:

- **OSM_NordKivu:** Open Street Map data points with settlement names in the Democratic Republic of Congo Nord Kivu province (https://download.geofabrik.de/africa/congo-democratic-republic.html)
- **popDensity_NordKivu:** High Resolution Settlement Layer in DRC (filtered for the Nord Kivu province) provided by Facebook Connectivity Lab and Center for International Earth Science Information Network - CIESIN - Columbia University – from 2015. To create these high-resolution maps, Facebook has used machine learning techniques to identify buildings from commercially available satellite images. Then they have overlaid general population estimates (from Columbia University) based on publicly available census data and other population statistics. The dataset includes 4.033.220 points which represent a 30 m grid with a value which corresponds to a population number estimate. (https://data.humdata.org/dataset/highresolutionpopulationdensitymaps-cod).



3. Population density-based clustering

In this step of the process we aimed to create clusters from the Facebook data points. These clusters represent population agglomerates from which it will be possible to create an estimate of the population number.

Density-Based Clustering is an unsupervised learning method which identifies distinctive groups or clusters in data, based on the idea that a cluster in a data space is a region of high point density, separated from other clusters by regions of low point density.

This ML algorithm is used in Python as follows:

```
sklearn.cluster.DBSCAN(eps, min_samples).fit(X, sample_weight)
```

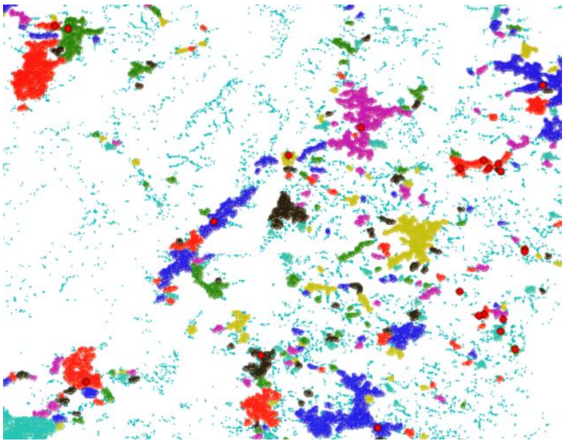
with eps=0.0009 (maximum search distance), min_samples=12 (minimum number of samples - or total weight, in a neighborhood for a point to be considered as a core point), and sample_weight equal to a vector with the same length as X, with a weight value for each sample (1 for regular points and 12 for core points).

The sample_weight vector is crucial to the success of this experiment. The cores (with weight=12) are the closest PopDensity_NordKivu points to the OSM_NordKivu point. This means that we are forcing our OSM settlements to be core points and thus forcing an agglomeration of points around this settlements. This is done in order to make sure the population agglomerates are generated around the right location – OSM settlement.

The parameters reported above result from an extensive trial-and-error approach which included many variations in these parameters.

4. Joining OSM and cluster data

After creating clusters from the PopDensity data, each core point labeled with a cluster number. This cluster numbers are used to distinguish clusters from each other. This distinction is then used to calculate the total population number for each cluster from the population number of each PopDensity point.



5. Potential of Population Estimate for each settlement

The processes described ultimately add value to our data. An **estimate of the population number** from 2015 data was associated with the settlements represented in the OSM dataset

This prediction could be especially significant in places with less data on other platforms, which is the case of the majority of villages and towns. It is often difficult to find a population estimate for these places as they do not belong to the list of major cities in the country. Crossing information between the OSM and Facebook Connectivity Lab datasets has good potential to shine light on this matter and thus add value to OSM datasets.